Using Multi-level Models to Assess Data From an Intelligent Tutoring System

Danielle S. McNamara & Jennifer L. Weston

Presented at the Proceedings of the 6th International Conference on Educational Data Mining 2013

Using Multi-level Models to Assess Data From an Intelligent Tutoring System

Jennifer L. Weston
Department of Psychology and
Learning Sciences Institute
Arizona State University
Tempe, AZ
Jen.weston@asu.edu

Danielle S. McNamara
Department of Psychology and
Learning Sciences Institute
Arizona State University
Tempe, AZ
Danielle.mcnamara@asu.edu

ABSTRACT
Intelligent tutoring systems yield data with many properties that render it potentially ideal to examine using multi-level models (MLM). Repeated observations with dependencies may be optimally examined using MLM because it can account for deviations from normality. This paper examines the applicability of MLM to data from the intelligent tutoring system Writing-Pal using intraclass correlations. Further analyses were completed to assess the impact of individual differences on daily essay scores along with the differential impact of daily vs. mean attitudinal ratings.

Keywords

Multi-Level models, Writing, Intelligent Tutoring Systems

1. INTRODUCTION
With the advent of intelligent tutoring systems (ITSs), the amount and complexity of data available to researchers has increased exponentially. ITSs provide the opportunity for repeated administration of assessments and, in some cases, ease of scoring that data. Though most tutoring systems provide multiple assessments of student progress (i.e., multiple text responses or worked problems), many researchers assess performance using pretest-posttest differences or repeated measures analyses, potentially missing out on rich data collected between these two end points.

When a student produces multiple responses, dependency arises in the data, thus violating central assumptions underlying both regression and ANOVA. Dependency, measured using intraclass correlations (ICC), is a pervasive problem in educational data, ranging from less problematic (a group of students within schools) to highly problematic (observations within individuals) [1]. Even when 5% of the variation in a data set is due to nested structure, (i.e.; dependency) it is advisable to assess differences at the highest cluster level.

The Writing Pal (W-Pal, [2]) is an ITS that provides writing strategy instruction to high school and entering college students. This system teaches writing strategies that encompass the entire writing process from prewriting through revision. Students have the opportunity to watch lesson videos, practice individual strategies within educational mini-games, and write and receive feedback on timed, prompt-based (SAT-style) essays.

In addition to providing instruction, W-Pal affords students the opportunity to practice writing and receive feedback on their essays. Students write prompt-based persuasive essays within an essay writing module. Essays are scored using an algorithm trained on a large corpus of SAT-style essays [3]. In this paper, we evaluate the applicability of multi-level modeling (MLM) for ITS data. Specifically, we examine the level and impact of dependency in the data. We examine a means-as-outcomes model assessing the impact of individual differences on daily essay scores. In addition, we examine a contextual effects model that assesses the differential impact of daily and mean ratings of attitudinal measures.

2. METHODS
Sixty-five high school students from a large urban southwestern city participated for payment in a lab based study to assess the effectiveness of W-Pal. All participants were recruited from the community. The study compared two versions of the W-Pal system: the full W-Pal system, and a version including only Essay Practice. In the W-Pal condition, students had access to the entire W-Pal system, whereas those in the Essay Practice condition only interacted with the essay practice function. These conditions were designed to control for time-on-task.

This study consisted of 10 sessions along with a home survey, which participants completed prior to attending their sessions. The home survey included basic demographics and measures of writing habits. The first session was a pretest session during which participants completed a pretest essay and prior knowledge assessments.

Participants in all conditions began sessions 2-9 by filling out a survey about their previous session and current mood, and then completed a SAT-style practice essay. Based on students' randomly assigned condition, some students interacted with all of W-Pal (n= 33), while others interacted with the Essay Practice module in W-Pal (n=32). Participants were given a maximum of 25-minutes to complete their essay. They then received feedback and were given an additional 10-minutes to revise their essays. Students in the W-Pal condition then completed an assigned lesson and game based practice. Students in the Essay condition completed a second SAT-style essay, also revising this essay.

During the final session, students completed a posttest, which was the same for all participants regardless of condition. For the current paper, only the essay scores, pretest, and attitudinal measures will be considered.

2.1 Measures

2.1.1 Essays
Depending on condition, participants wrote either 8 or 16 practice essays with feedback, ,and a pretest and posttest essay without feedback. The essay prompts were adapted from SAT writing assessments and scored on a 1-6 scale using the W-Pal algorithm validated by Crossley and colleagues [3]. This algorithm displays sufficient accuracy (exact agreement of 55% and adjacent

agreement of 92%). The present analyses focus on the eight practice essays with common prompts for both conditions (i.e., the first essay written in the Essay condition). The pretest essay provides a measure of prior writing ability.

2.1.2 Individual Difference Measures

A variety of individual difference measures were administered to assess the impact of these characteristics on essay quality. In the present study, we focus on the measures of self-efficacy, prior reading ability, and motivation. Self-efficacy was measured using the Writing Attitudes and Strategies Self-Report Inventory (WASSI, [4]). Prior reading ability was assessed using the Gates MacGinitie Reading Test (GMRT Ed.3; level 10/12, form S). Motivation was measured using a daily and posttest survey with questions about participants' moods and previous and anticipated interactions with W-Pal.

3. RESULTS

3.1 Applicability of Multilevel Models

A series of unconditional models for all level-1 variables were estimated. The variance estimates from these analyses were used to compute intraclass correlations (ICCs). The ICC for daily essay scores was ICC =.47, suggesting that 47% of the variance in essay score can be attributed to the individual. For daily survey items, these values ranged from .37 - .98, suggesting that a significant portion of the variance for all of the daily survey items can be attributed to the individual.

3.2 Means-as-Outcomes Model

We estimated a means-as-outcomes model in which we used a number of level-2 variables to predict daily essay score. Variables were selected based on prior research on writing and included prior writing ability, reading ability (GMRT), writing self-efficacy, and condition. This model assesses the impact of each prior ability measure on average daily essay score holding all others constant.

A likelihood-ratio test was completed to assess the explanatory power of the level-2 variables. The Likelihood-ratio test was significant $\chi^2(4) = 46.21$, $p < .001$, suggesting that the MLM is superior to a model not containing these variables. The Bayesian Information Criterion (BIC) was also examined. The results from the BIC values mirrored the results found using the likelihood ratio test [5] Additionally, the inclusion of these five variables reduced the between cluster variation by 63%. All predictors had a significant impact on daily essay scores (prior writing ability B = .211, Prior Reading Ability B = .034, Self-Efficacy B = .023, and Condition B = .014)

3.3 Contextual Effects Model

An additional model was estimated using the daily survey data to predict daily essay scores. To investigate the possibility of contextual effects (differential effects at level-1 and level-2), we also included the cluster (person) means as level-2 predictors.

A Wald test of the 10 level-2 coefficients was statistically significant, $F(10, 23) = 23.943$, $p = .007$, indicating that the set of contextual effects improved the fit of the model. Further univariate tests indicated that competitiveness ($\gamma 1$), feelings of frustration ($\gamma 2$), and self-assessments of improvement ($\gamma 3$) exerted significant contextual effects, $\gamma 1 = .102$, $p = .003$; $\gamma 2 = -.070$, $p = .020$; $\gamma 3 = -.261$, $p = .049$; the contextual effect for mood ($\gamma 4$) was marginally significant, $\gamma 4 = .373$, $p = .061$. The signs and magnitude of the level-2 regressions (daily survey means predicting daily essay mean) were stronger than the level-1 predictors; however, the effects of sustained levels of certain feelings about the system (e.g., frustration) seemed to be more complex, warranting further investigation.

4. DISCUSSION

The data examined in this study exhibit high levels of dependency, rendering it ideal for multi-level modeling. The ICC values for the repeated assessments in W-Pal range from .37 - .98, exceeding appropriate values for using regression and analysis of variance. By using a means-as-outcomes model, we were able to account for 63% of the variance due to the cluster (student). The results suggest that there is an advantage for those interacting with the complete W-Pal system, additionally, individual differences were important predictors of average daily essay score.

The analysis using the contextual effects model showed that, for this data, daily and mean values for attitudinal survey items had differential effects on essay scores. For instance, while daily enjoyment has a negative relationship with daily essay score, the participant's average level of enjoyment had a positive relationship with average essay scores.

Further work will be completed to combine these models and to investigate the utility of using random slopes for the level-1 variables. Interactions will also be investigated further. Overall, the data from W-Pal is ideal for using MLM for assessment.

5. ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080589 to Arizona State University.

6. REFERENCES

[1] Hedges, L. V., and Hedberg, E. C.(2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis.* 29, 1. 60-87.

[2] Roscoe, R. D., and McNamara, D.S. (in press). Writing pal: intelligent tutoring of writing strategies in the high school classroom, *Journal of Educational Psychology.*

[3] Crossley, S. A., Roscoe, R., and McNamara, D. S. (in press). Predicting human scores of essay quality using computational indices of linguistic and textual features. Proceedings of the 15th International Conference on Artificial Intelligence in Education. Auckland, New Zealand: AIED.

[4] Weston, J. L., Roscoe, R., Floyd, R. G., and McNamara, D. S. (2013, May). The WASSI (Writing Attitudes and Strategies Self-Report Inventory): Reliability and validity of a new self-report writing inventory. Poster presented at the 2013 Annual Meeting of the American Educational Research Association, San Francisco, CA

[5] Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systemic Biology,53*, 5. 793-808.